Population Size Estimation Using Individual Level Mixture Models

Daniel Manrique-Vallier*and Stepthen E. Fienberg †‡

April 16, 2008

Abstract

We revisit the heterogeneous closed population multiple recapture problem, modeling individual-level heterogeneity using the Grade of Membership model (Woodbury et al., 1978). This strategy allows us to postulate the existence of homogeneous latent "ideal" or "pure" classes within the population, and construct a soft clustering of the individuals, where each one is allowed *partial* or *mixed membership* in all of these classes. We propose a full hierarchical Bayes specification and a MCMC algorithm to obtain samples from the posterior distribution. We apply the method to simulated data and to three real life examples.

Keywords: Capture-recapture, Dependence, Grade of Membership models, Heterogeneity, Hierarchical Bayes, BIC, Log-linear models, Rasch model.

1 Introduction

From the early work of Petersen (1896) to counting in fish populations (but see Goudie and Goudie (2007)) through the 1960s, e.g., see Cormack (1968), most of the focus in the literature on capture recapture approaches to the estimation of the size of closed biological populations focused on the independence of captures or lists and homogeneous capture probabilities within lists. Attention to comparable problems in non-animal populations came later but still had a similar focus and used by and large similar models. It was only with the emergence of models for multivariate categorical data in the 1960s and a shift to other types of applications that different proposals emerged for incorporating departures from this basic homogeneous independence structure in the form of heterogeneity of capture probabilities among individuals (units),

*Corresponding author: e-mail: dmanriqu@stat.cmu.edu Department of Statistics. Carnegie Mellon University. Pittsburgh, PA 15213, USA

†e-mail: fienberg@stat.cmu.edu. Department of Statistics and Machine Learning Department. Carnegie Mellon University. Pittsburgh, PA 15213, USA

‡This research was supported in part by the National Institutes of Health under Grant No. R01 AG023141-01 to Carnegie Mellon University. We are also grateful with Matthew S. Johnson for sharing his implementation of the MCMC sampler to fit the Bayesian Rasch model.

e.g., see Sanathanan (1972, 1973) and/or associations between or among lists, see Fienberg (1972); Bishop et al. (1975). Simultaneous consideration of both types of departures appeared in the literature only in the 1990s, although even then, most approaches to heterogeneity assumed a very specific form, such as that associated with the Rasch model (Rasch, 1980) from item response theory in psychometrics, e.g., see Darroch et al. (1993); Agresti (1994); International Working Group for Disease Monitoring and Forecasting (1995a,b); Fienberg et al. (1999). Similar models were introduced in the animal multiple recapture literature but with different notation, e.g., see Norris and Pollock (1996). In some applications such approaches have proven quite successful but in others we appear to require a more flexible form of latent variable model.

In this paper we propose an individual mixture approach to dependence in multiple recapture problems based on the Grade of Membership (GoM) model introduced by Woodbury et al. (1978) and developed from a hierarchical Bayesian latent variable perspective by Erosheva (2002). For full details and an application to the analysis of disability data, see Erosheva et al. (2007). This mixture model has a different interpretation than most others in the statistical literature and especially those used in a multiple recapture context, c.f. Pledger (2000, 2005).

In section 2, we describe the GoM model and then explain how it can be adapted for use in the context of multiple recapture estimation. In section 3, we outline the details regarding estimation using Monte Carlo Markov chain methodology. Then, in section 4, we apply the new model to a series of examples drawn from the literature and compare the resulting estimates to those derivable from log-linear models and Rasch models for heterogeneity. We conclude with a discussion of where we see the new model fitting into the array of existing approaches as a practical tool for applied researchers.

2 Modeling heterogeneity using the Grade of Membership model

2.1 The Grade of Membership model

The Grade of Membership (GoM) model has been used to get a low dimensional representation for high-dimensional categorical data where we think of each individual as a weighted combination of a small number of "idealized" individuals or "pure types" within the population of interest.

We assume the existence of a specific number, K, of such "extreme classes" or "pure types". Suppose that there are N individuals on the population. For the ith individual, for $i \in \{1, \ldots, N\}$, we

associate a J-dimensional binary vector of manifest variables $x_i = (x_{i1}, ...x_{iJ})$. For any individual that is a *full member* of the kth extreme class (i.e. an "ideal" individual of the kth class), we assume that the probability of a positive response in the jth entry of the manifest variables vector is the same, i.e., $\Pr(X_{ij} = 1|i\text{th individual in } k\text{th class}) = \lambda_{jk}$.

We associate each individual with its own K-dimensional "membership vector" $g_i = (g_{i1}, g_{i2}, ..., g_{iK})$ representing how much of a member of each class this particular individual is ($g_{ik} > 0$, $\sum_{k=1}^{K} g_{ik} = 1$). We operationalize the idea of "partial membership" by setting the distribution of each manifest variable given the membership vector as

$$p(x_{ij}|g_i) = \sum_{k=1}^{K} g_{ijk} \lambda_{jk}^{x_{ij}} \left(1 - \lambda_{jk}\right)^{1-x_{ij}}.$$

In what follows, we use $p(\cdot)$ to denote indistinctly the probability density function or the probability mass function of the argument, as needed. We further assume that the item responses j are conditionally independent given membership vectors. This condition, sometimes referred as *latent conditional independence* or *local independence* (Holland and Rosenbaum, 1986), expresses the idea that the membership vector g completely explains the dependence structure between the J binary manifest variables. By making this assumption, we get

$$p(x_i|g_i) = \prod_{j=1}^{J} \sum_{k=1}^{K} g_{ijk} \lambda_{jk}^{x_{ij}} \left(1 - \lambda_{jk}\right)^{1-x_{ij}}. \tag{1}$$

If we further assume that for each individual, the (unobserved) membership vectors g_i are drawn independently from a common distribution G_α, with support on the $K-1$ dimensional probability simplex, Δ, we can construct the observed likelihood

$$p(x_i) = \int_{\Delta} \prod_{j=1}^{J} \sum_{k=1}^{K} g_{ijk} \lambda_{jk}^{x_{ij}} \left(1 - \lambda_{jk}\right)^{1-x_{ij}} G_\alpha(dg). \tag{2}$$

The net effect is that we are representing a collection of 2^J counts in terms individual mixtures of K pure types, thus producing a very substantial reduction in dimensionality. The only problem is we still have to determine the pure types for a given value of K.

Erosheva et al. (2007) have shown that the GoM model in (1) admits an augmented data representation

that leads to the augmented data likelihood

$$p(x,z|\lambda,g)=\prod_{i=1}^{N}\prod_{j=1}^{J}\prod_{k=1}^{K}\left(g_{ik}\lambda_{jk}^{x_{ij}}\left(1-\lambda_{jk}\right)^{1-x_{ij}}\right)^{z_{ijk}}, \tag{3}$$

with $z_i=(z_{i1},...,z_{iJ})\in Z=\{1,2...,K\}^J$ and $z_{ijk}=I(z_{ij}=k)$. From here the observed data likelihood can be easily obtained through marginalization

$$p(x|\alpha,\lambda)=\prod_{i=1}^{N}\int_{\Delta}\sum_{z\in Z}\prod_{j=1}^{J}\prod_{k=1}^{K}\left(g_{k}\lambda_{jk}^{x_{ij}}\left(1-\lambda_{jk}\right)^{1-x_{ij}}\right)^{z_{jk}}G_{\alpha}(dg). \tag{4}$$

The representation in Eq. (3) is particularly important because we use it as the basis for the posterior sampling strategy we describe in Section 2.2.2.

2.2 Hierarchical Bayes formulation of multiple recapture estimation of a closed population total using the Grade of Membership model

2.2.1 Basic setup

Consider a closed population of N individuals, with N unknown. From this population we perform J capture events. These "capture events" can be understood as the physical capture of animals in ecology applications or as the elaboration of lists in epidemiological applications. We represent the complete capture history of the ith individual ($i=1,...,N$) by a vector $x_i=(x_{i1},x_{i2},...,x_{iJ})$ with $x_{ij}=1$ if the i-th individual was captured in the jth capture event and 0 otherwise. We assume that from the N individuals present in the population, we are only able to capture or observe n of them ($n\leq N$), where all capture history vectors of the form $x_i=(0,0,...,0)$ are missing for all the cases. We arrange the capture history so that $x_i=(0,0,...,0)$ for all $i=n+1,...,N$.

2.2.2 Model specification

Using Eq. (3) for the likelihood of the GoM model and understanding N as a parameter, we get that the full joint posterior for the complete data will be,

$$p(N,\alpha,\lambda,g|x,z)\propto\binom{N}{n}p\left(N,\alpha,\lambda,g\right)\prod_{i=1}^{N}\prod_{j=1}^{J}\prod_{k=1}^{K}\left(g_{ik}\lambda_{jk}^{x_{ij}}\left(1-\lambda_{jk}\right)^{1-x_{ij}}\right)^{z_{ijk}}I_{\{n+1,n+2,...\}}(N), \tag{5}$$

where $p(N, \alpha, \lambda, g)$ is the joint prior distribution of (N, α, λ, g). If we further assume that N, α and λ are a priori independent and that given α, g_i are independent for all i, the above expression simplifies to

$$p(N, \alpha, \lambda, g|x, z) \propto \binom{N}{n} p(N)p(\alpha)p(\lambda) \prod_{i=1}^{N} p\left(g_i|\alpha\right) \prod_{j=1}^{J} \prod_{k=1}^{K} \left(g_{ik} \lambda_{jk}^{x_{ij}} \left(1 - \lambda_{jk}\right)^{1-x_{ij}}\right)^{z_{ijk}} I_{\{n+1,...\}}(N). \quad (6)$$

We complete the specification by setting the hyper priors

$$\lambda_{jk} \overset{iid.}{\sim} Beta(\eta_1, \eta_2), \quad (7)$$

$$g_i|\alpha \overset{iid.}{\sim} Dirichlet(\alpha_1, \alpha_2, ..., \alpha_K), \quad (8)$$

$$p(N) \propto \frac{(N-l)!}{N!} I_{\{[l, N_{max}] \cap \mathbb{N}\}}(N) \quad \text{for some } l < N_{max} \text{ and } N_{max} \text{ large enough.} \quad (9)$$

The selected prior for N in Eq. (9) has the advantage of leading to proper and tractable expressions for the posterior while including many relevant cases, such as the truncated uniform ($l = 0$) and the truncated Jeffrey's ($l = 1$) (Fienberg et al., 1999). The upper truncation limit, N_{max}, is included for computational convenience and because its inclusion guaranties a proper posterior. Also, we note that, although the prior distribution for N in principle allows it, we will not want N to take values below the observed count, n. This will be automatically enforced during the posterior computation, because the likelihood in Eq. (6) will in fact truncate and re normalize the posterior according to the data, making values of N smaller than n to have probability zero.

We follow Erosheva et al. (2007) by reparameterizing $\alpha = (\alpha_0 \cdot \xi_1, ..., \alpha_0 \cdot \xi_K)$ with $\alpha_0 > 0$ and $\sum_k \xi_k = 1$ and setting the hyper priors

$$\alpha_0 \sim Gamma(\tau, \beta), \quad (10)$$

$$\xi \sim Dirichlet(1, ..., 1), \quad (11)$$

where β is the inverse scale parameter. Under this reparametrization we can interpret ξ_k as the proportion of the captures that belongs to the kth mixture component and $\alpha_0 > 0$, as a parameter governing the spread of the membership distribution.

3 Bayesian posterior estimation using Monte Carlo Markov chain

If N were known, generating samples from the posterior distribution in Eq. (6) would be a straightforward exercise, directly applying the algorithm developed in Erosheva et al. (2007). Here we present an adaptation of this algorithm to carry out our computations.

The main difficulty in constructing a MCMC sampler for this case is the fact that, as Basu and Ebrahimi (2001) and Fienberg et al. (1999) point out, N determines the length of the vectors g and z, and N is itself a parameter that we have to estimate. For this reason we consider the partitions $g = (g_1, g_2)$, $z = (z_1, z_2)$, $x = (x_1, x_2)$ such that the first part (g_1, z_1, x_1) corresponds to all individuals who are present in the sample $(i = 1, ..., n)$, and the second part (g_2, z_2, x_2) corresponds to all the individuals who were not captured $(i = n + 1, ..., N)$.

From the specification of the model in Eq. (6), we can readily write down the full conditional posteriors for λ, z_1 and g_1, and obtain the $(m + 1)$-th draw for a Gibbs sampler,

- z_1: for $i = 1...n$ and $j = 1...J$ draw

$$z_{ij}^{(m+1)} \overset{indep}{\sim} Discrete_{(1,...,J)}(p_1, ..., p_K), \tag{12}$$

  with $p_k = g_{ik}\lambda_{jk}^{x_{ij}}(1 - \lambda_{jk})^{1-x_{ij}}$.

- λ: for $j = 1...J$ and $k = 1...K$ draw

$$\lambda_{jk}^{(m+1)} \overset{indep}{\sim} Beta\left(\eta_1 + \sum_{i=1}^{N} x_{ik}z_{ijk}, \eta_2 + \sum_{i=1}^{N}(1 - x_{ij})\, z_{ijk}\right). \tag{13}$$

- g_1: for $i = 1...n$ draw

$$g_i^{(m+1)} \overset{indep}{\sim} Dirichlet\left(\alpha_1 + z_{i+1}, ..., \alpha_K + z_{i+K}\right), \tag{14}$$

  where $z_{i+k} = \sum_{j=1}^{J} z_{ijk}$.

The full conditional distribution for α does not have any recognizable form. Thus we use a Metropolis-Hastings within Gibbs step (Robert and Casella, 2004). Erosheva (2002) proposes a two-stage sampling

strategy, obtaining samples for α_0 and for ξ through the use of two separate Metropolis-Hastings steps. We have achieved better results treating the vector α as a whole using the following log scale Gaussian random walk Metropolis-Hastings step:

1. (Proposal step) Obtain a sample for $\alpha^* = (\alpha_1^*, ..., \alpha_K^*)$ from

$$\alpha_k^* \overset{indep.}{\sim} lognormal(\alpha_k^{(m)}, \sigma^2),$$

for $k = 1, ..., K$. The parameter σ^2 is a tuning parameter that we have to optimize to achieve a good balance between acceptance and exploration of the support (Robert and Casella, 2004). Note that the distribution of α^* depends explicitly on the current value of α.

2. (Acceptance step) compute

$$r = \min\left\{1, \left(\prod_{k=1}^{K}\frac{\alpha_k^*}{\alpha_k}\right)\left(\frac{\alpha_0^*}{\alpha_0}\right)^{\tau-1} e^{-\beta(\alpha_0^*-\alpha_0)}\left[\frac{\Gamma(\alpha_0^*)}{\Gamma(\alpha_0)}\prod_{k=1}^{K}\frac{\Gamma(\alpha_k^*)}{\Gamma(\alpha_k)}\right]^N \prod_{k=1}^{K}\left(\prod_{i=1}^{N} g_{ik}\right)^{\alpha_k^*-\alpha_k}\right\},$$

and make

$$\alpha^{(m+1)} = \begin{cases} \alpha^* & \text{with probability } r. \\ \alpha^{(m)} & \text{with probability } 1-r. \end{cases}$$

To obtain a full conditional posterior of (N, z_2, g_2), we first note that it only depends on (α, λ). Then we apply the well known factorization

$$\begin{aligned} p(N, z_2, g_2|...) &= p(N, z_2, g_2|\alpha, \lambda) \\ &= p(N|\alpha, \lambda)p(z_2|N, \alpha, \lambda)p(g_2|z_2, N, \alpha, \lambda), \end{aligned}$$

and construct the relatively simpler samplers for the incomplete conditionals $p(N|\alpha, \lambda)$ and $p(z_2|N, \alpha, \lambda)$ by integration,

$$\begin{aligned} p(N|\alpha, \lambda) &= \int p(N, z_2, g_2|\alpha, \lambda)d(g_2, z_2) \\ &\propto \binom{N}{n} p(N)\left[\frac{\Gamma(\alpha_+)}{\Gamma(\alpha_+ + J)}\sum_{z\in Z}\left(\prod_{k=1}^{K}\frac{\Gamma(\alpha_k + z_{+k})}{\Gamma(\alpha_k)}\right)\prod_{j=1}^{J}\left(1-\lambda_{jz_j}\right)\right]^{N-n} \end{aligned} \tag{15}$$

for N integer such that $N \geq n$; and

$$
\begin{aligned}
p(z_2|N,\alpha,\lambda) &= \int p(z_2,g_2|N,\alpha,\lambda)dg_2 \\
&\propto \prod_{i=n+1}^{N}\left[\left(\prod_{k=1}^{K}\Gamma\left(\alpha_k+z_{i+k}\right)\right)\prod_{j=1}^{J}\left(1-\lambda_{jz_{ij}}\right)\right] \qquad (16)
\end{aligned}
$$

where $Z=\{1,2,...,K\}^J$, $z_{jk}=I_{\{z_j=k\}}$, $z_{+k}=\sum_{j=1}^{J}z_{jk}$ and $\alpha_+=\sum_{k=1}^{K}\alpha_k$.

Finally we note that $p(g_2|z_2,N,\alpha,\lambda)$ is just the full conditional distribution of g_2,

$$
p(g_2|N,z_2,\alpha,\lambda) \propto \prod_{i=n+1}^{N}\prod_{k=1}^{K} g_{ik}^{\alpha_k+z_{i+k}-1}. \qquad (17)
$$

One single joint sample $(N,z_2,g_2)^{(m+1)}$ from the full conditional distribution $p(N,z_2,g_2|...)$ will be then constructed sampling sequentially from the distributions in Eqs. (15), (16) and (17). Note also that for our selection of priors, Eq. (15) is truncated negative binomial on $N-n$ (Fienberg et al., 1999), and that a sample from Eq. (17) can be constructed as $N-n$ independent samples from a $Dirichlet(\alpha_1+z_{i+1},...,\alpha_K+z_{i+K})$ distribution, just as in Eq. (14).

4 Examples

We have applied our new model and method to a number of examples from the literature; here we report on four examples, the first of which is based on a simulation from the GoM model in Eq. (1). The remaining three examples are based on actual data drawn from a series of applications with increasing numbers of lists: diabetes patients in Casale Monferrato, Italy (Bruno et al., 1994); children suffering a specific congenital anomaly in Massachusetts (Wittes et al., 1974; Fienberg, 1972); and killings and disappearances due to political violence in the District of Chungui, Peru (Ball et al., 2003).

For all the examples we have fitted the GoM model using the MCMC algorithm from Section 2.2.2 for

different values of K using the diffuse priors

$$\begin{aligned} \lambda_{jk} &\overset{iid}{\sim} Beta(1,1), \\ \alpha_0 &\sim Gamma(1,2), \\ \xi &\sim Dirichlet_K(1,...,1), \\ p(N) &\propto I_{\{N<10^5 \cap \mathbb{N}\}}. \end{aligned}$$

For comparison, we have also fitted independence and more complex log-linear models (Fienberg, 1972) and the Bayesian Rasch model from Fienberg et al. (1999). The selection of the "best" log-linear model has been performed using stepwise search based on the Bayesian Information Criterion (BIC). The stepwise procedure was implemented similarly to a regression model search: starting from the independence model, we added new interaction terms, one at a time, so that at each step we obtained the biggest improvement in the BIC score computed over the observed data, until no further improvement was possible.

For each example we present the point estimate ($\hat{N}$) and a 95% interval. For the log-linear models the interval is an equal tail 95% bootstrap confidence interval (Efron and Tibshirani, 1993) and for the Bayesian models (GoM and Rasch) it is an equal tail 95% credible interval constructed from the posterior sample. The point estimates presented for the Bayesian models are the posterior modes.

The chains we obtained fitting the GoM models were reasonably well behaved, converging to the stationary distribution after a burn in period of 10000 iterations for $K = 2$ and 20000 for $K = 3$ for all the cases except for the Perú data, where we required longer runs of 30000 iterations for $K = 3$, 55000 for $K = 4$ and 30000 for $K = 5$. In all cases we took a total of 100000 samples and sub sampled the resulting chains retaining one sample every t samples and discarded the rest, with t ranging from 10 to 40.

				List-4			
				Yes		No	
				List-3		List-3	
				Yes	No	Yes	No
List-2	Yes	List-1	Yes	161	40	48	10
			No	494	88	41	1
	No	List-1	Yes	73	11	5	2
			No	123	193	33	**677**

Table 1: Simulated data ($N = 2000$, observed $n = 1323$)

Model	df	Deviance	$\hat{N}$	95% Interval
Independence	10	83.06	2425	[2269, 2586]
Log-linear-BIC ([3][12][24])	8	12.46	1994	[1878, 2140]
Bayesian Rasch	—	—	2773	[2511, 3252]
Bayesian GoM ($K = 2$)	—	—	2001	[1881, 2235]
Bayesian GoM ($K = 3$)	—	—	2060	[1909, 2327]

Table 2: Estimates for the simulated data ($N = 2000$, observed $n = 1323$)

4.1 Simulated data ($J = 4$)

We simulated 2000 draws from the GoM model (1) with $J = 4$ lists, $K = 3$ extreme profiles, membership vectors $g_i \overset{iid}{\sim} Dirichlet\,[0.8 \cdot (0.1, 0.5, 0.4)]$ and extreme profile capture probabilities

$$[\lambda_{jk}] = \begin{bmatrix} 0.5 & 0.09 & 0.2 \\ 0.1 & 0.7 & 0.01 \\ 0.29 & 0.14 & 0.1 \\ 0.05 & 0.035 & 0.35 \end{bmatrix}$$

The resulting data is summarized in the 2^4 cross classification in Table 1. Our goal is to predict the total (known to be $N = 2000$) using only the counts on the cells other than $(0, 0, 0, 0)$.

In Table 2 we present a summary of our results. As expected, the independence model fits poorly (with a deviance of $G^2 = 83.06$ on 10 df.), and overestimates the total. The BIC-based stepwise search in the space of hierarchical log-linear models selects the model [3][12][24], which provides both a very good fit and estimate of N. In contrast, the Bayesian Rasch model misses the true value, grossly overestimating it. This appears to be due to the inability of the Rasch model to accommodate negative dependence. Negative dependence may also explain why the independence model also overestimates the true value of N.

				Clinics(4)			
				Yes		No	
				Hospitals(3)		Hospitals(3)	
				Yes	No	Yes	No
Prescriptions(2)	Yes	Reimbursements(1)	Yes	58	46	14	8
			No	157	650	20	182
	No	Reimbursements(1)	Yes	18	12	7	10
			No	104	709	74	**??**

Table 3: Diabetes data

For fitting the GoM models we have applied the MCMC algorithm in 2.2.2, taking $100,000$ samples and discarding the first $20,000$ as a burn in period. The resulting samples are well behaved, leading to estimates close to the true value. Interestingly both two GoM ($K = 2$ and $K = 3$) models lead to good estimates for N, even when the data was actually generated using $K = 3$.

During the creation of this example we have tried different combinations of parameters. In most cases, the parameter N was correctly estimated by the GoM models, as might be expected, and also by the log-linear-BIC. In some other cases cases though, only the GoM model produced a good estimate. However in many of these difficult cases we were unable to consistently replicate the simulation and obtain good estimates every time.

4.2 Diabetes data

Bruno et al. (1994) used multiple recapture data from four sources to estimate the prevalence of diabetes mellitus in Casale Monferrato, Italy on October 1st, 1988. They observed a total number of $n = 2069$ known cases and applying log-linear modeling to obtain an estimate of $\hat{N} = 2,771$ for the total number of cases. When they first stratified the data and then fit the same log-linear model, their estimate was $\hat{N} = 2,586$. These data have been re analyzed a number of times, e.g., International Working Group for Disease Monitoring and Forecasting (1995a); Fienberg et al. (1999). Table 3 shows the cross classification of the counts for the four different sources used to identify the known diabetes cases as reported by Fienberg et al. (1999).

Table 4 shows a summary of our estimates using this dataset. As in the previous example, we infer from the deviance that the fit of the independence model is poor. Both the Rasch and the "best" log-linear approaches give estimates very close to the ones reported by Bruno et al. (1994), International Working Group for Disease Monitoring and Forecasting (1995a), i.e., $\hat{N} = 2771$. The log-linear-BIC model also provides a

Model	df	Deviance	$\hat{N}$	95% Interval
Independence	10	217.48	2251	[2228, 2275]
Log-linear-BIC ([12][23][24][34])	5	7.617	2771	[2543, 3097]
Bayesian Rasch	—	—	2680	[2549, 2934]
Bayesian GoM ($K = 2$)	—	—	2293	[2250, 2347]
Bayesian GoM ($K = 3$)	—	—	2640	[2483, 3482]

Table 4: Estimates for the diabetes data

				List-I							
				Yes				No			
				List-II				List-II			
				Yes		No		Yes		No	
				List-III		List-III		List-III		List-III	
				Yes	No	Yes	No	Yes	No	Yes	No
List-IV	Yes	List-V	Yes	2	8	3	5	0	23	0	30
			No	5	25	1	22	3	37	2	97
	No	List-V	Yes	2	18	5	36	0	34	3	83
			No	1	19	4	27	1	37	4	**??**

Table 5: Congenital anomalies data

very good fit to the observed data ($G^2 = 7.617$ on 5 df.)

The GoM model with $K = 2$ gives an estimate of N very close to the independence model, which we believe is a bad estimate; however, when we increase the number of extreme profiles to $K = 3$, the GoM model again estimates N in the vicinity of 2,700, as does the best log-linear model and the Rasch model.

4.3 Congenital anomalies data ($J = 5$)

Fienberg (1972) reports on a 5-list multiple recapture dataset originally analyzed by Wittes et al. (1974) regarding positive diagnostics of a specific congenital anomaly in children in Massachusetts. The dataset presents five available sources of positive diagnostics, labeled List-I to List-V, and a total of $n = 537$ known cases. Fienberg (1972) reports a best estimate of $\hat{N} = 634$ using the loglinear model [12][13][24][45].

In Table 5 we reproduce the 2^5 cross classification table as reported by Fienberg (1972). Table 6 shows our estimates for this dataset. As in all previous cases, independence seems to be a bad fit ($G^2 = 93.45$ on 25 df.). The best log-linear model chosen by a stepwise application of BIC is slightly simpler than the one proposed by Fienberg (1972), yielding to the slightly lower estimate $\hat{N} = 620$ ($G^2 = 30.56$ on 22 df).

Model	df	Deviance	$\hat{N}$	95% Interval
Independence	25	93.45	639	[620, 659]
Log-linear-BIC ([12][13][45])	22	30.56	620	[600,644]
Bayesian Rasch	—	—	712	[661, 824]
Bayesian GoM ($K = 2$)	—	—	605	[584, 636]
Bayesian GoM ($K = 3$)	—	—	622	[596, 706]
Bayesian GoM ($K = 4$)	—	—	620	[590, 717]

Table 6: Estimates for the congenital anomalies data

All the estimates from the GoM model are close to one another and give similar 95% intervals, suggesting that $K = 2$ might offer sufficient heterogeneity to represent that which is present in this sample. As in the previous cases, the estimates obtained from fitting the GoM models are very close to the estimate obtained from the log-linear-BIC. The Rasch model, however, gives an estimate well above the rest. This suggess again that the Rasch model is producing an overestimate due to to its inability to capture negative dependence.

4.4 Killings in Chungui ($J = 6$)

Ball et al. (2003) in a report published as part of the Peruvian Truth and Reconciliation Commission Final Report (Comisión de la Verdad y Reconciliación, 2003) analyzed multiple capture data to estimate the total number of people assassinated or disappeared due to political violence in Peru between 1980 and 2000. There were a total of six available lists documenting known cases of killings and disappearances, although one of them was restricted to only one geographical location and some others were quite small. Ball et al. (2003) estimations were constructed applying 3-list log-linear modeling to a total of 58 geographic strata, collapsing some of the sources based on homogeneity considerations.

Table 7 shows data corresponding to the district of Chungui, department of Ayacucho, where a total of $n = 1366$ unique individuals could be identified. This district corresponds a subset of stratum 25 in Ball et al. (2003)[1], and is the only geographical location for whom we have six simultaneously available lists. As we see in Table 7, the resulting cross classification is extremely sparse.

Table 8 summarizes our computations using different models. As in all previous cases, independence did not provide a good fit to the data ($G^2 = 709.6$ on 56 df.), but the log-linear-BIC approach yielded an

[1]Stratum 25 comprised the districts of Chungui and Luis Carranza, in the department of Ayacucho. The estimate from Ball et al. (2003) (as appears in Electronic Annex #12 Comisión de la Verdad y Reconciliación (2003)) for the whole stratum 25 is $\hat{N} = 2408$.

						List-VI							
						Yes				No			
						List-V				List-V			
						Yes		No		Yes		No	
						List-IV		List-IV		List-IV		List-IV	
						Yes	No	Yes	No	Yes	No	Yes	no
List-III	Yes	List-II	Yes	List-I	Yes	0	0	1	0	0	0	2	0
					No	0	0	0	0	0	0	0	3
			No	List-I	Yes	0	0	14	1	0	0	51	4
					No	0	0	1	1	0	0	1	8
	No	List-II	Yes	List-I	Yes	0	0	0	0	0	0	0	0
					No	0	0	2	0	0	0	0	1
			No	List-I	Yes	0	1	4	237	0	0	18	286
					No	0	0	0	727	0	0	3	??

Table 7: Killings in Chungui data

Model	df	Deviance	$\hat{N}$	95% Interval
Independence	56	709.6	1991	[1908, 2098]
Log-linear-BIC ([14][23][34][36])	52	56.63	2270	[2138, 2436]
Bayesian GoM ($K = 2$)	—	—	2244	[2086, 2437]
Bayesian GoM ($K = 3$)	—	—	2217	[2068, 2417]
Bayesian GoM ($K = 4$)	—	—	2231	[2064, 2390]
Bayesian GoM ($K = 5$)	—	—	2211	[2063, 2401]

Table 8: Estimates for killings in Chungui data. We were not able to fit the Bayesian Rasch model.

excellent fit ($G^2 = 6.63$ on 52 df.), giving an estimate of $\hat{N} = 2270$.

We were unable to fit the Rasch model in this example because all of our attempts to sample from the posterior distribution ended up with MCMC chains that were badly behaved, suggesting the inadequacy of the Rasch model to capture the specific type of heterogeneity present in this sample. In contrast, all GoM models behaved very well, giving estimates very close to the best log-linear model, although with a slightly wider 95% interval. The similarity of the estimates for $K = 2, 3, 4$ and 5 suggests that $K = 2$ would be enough to capture the heterogeneity structure present in this sample.

5 Discussion

In this article we have described a Bayesian version of the Grade of Membership model first introduced by Woodbury et al. (1978) and shown how we can use it to model individual level heterogeneity in multiple recapture contexts. We illustrated the methodology in a series of examples and compared the estimate from

our method with those from more traditional approaches.

The examples we analyzed share some interesting characteristics. First, in none of them does the independence assumption lead to good results. In fact, with the possible exception of the congenital anomalies data where there appears to be a canceling effect of positive and negative dependencies, assuming independence led to either over- or under-estimation of N, with deceptively tight confidence bounds. In contrast, when we allowed for more complex log-linear models and selected a best one by a stepwise search based on the BIC score, the results were consistently good. The use of the GoM model to estimate the population size always gave results comparable with the best log-linear models chosen by stepwise search based on BIC.

In our examples, low values of K appear to do a remarkably good job of capturing or approximating the heterogeneity structure present in the samples, leading to reasonable estimates of N in virtually all cases. In fact, in the examples and in other analyses we have performed, once for a given value of K we obtained an estimate comparable to that of the "best" log-linear model, all the subsequent GoM models, with higher values of K, yielded similar estimates. We believe this is due to the relatively small number of lists (4, 5, and 6) as compared with the large number of items in other applications of the GoM approach, e.g., $J = 16$ binary disability measures in Erosheva et al. (2007). Thus we believe one need not search for an "optimal" value of K for typical multiple recapture applications, and we can expect that $K = 2$ or $K = 3$ will suffice in almost all contexts.

A formal evaluation of the best number of extreme profiles, however, remains an open issue, especially for large numbers of lists. Erosheva et al. (2007) explored model selection with the Bayesian GoM model in the context of their analysis of disability data. In their analysis with simulated data they explored the performance of DIC, BICM, BIC, truncated Chi squared (χ^2_{tr}) and AICM indices and obtained good results with the last three. The application of these indices to evaluate the fit of the models based on the observed counts in the contingency tables looks like a natural starting point in our problem.

Our final example, estimating killings in an area of Peru, illustrates the applicability of different modeling approaches to sparse data situations, ones where most statistical analysts have in the past been reluctant to apply modern multiple recapture methods.

Finally, we note that Pledger et al. (2003) and others have explored the use of population-level mixture models and closed population estimators to the problem of open populations. We think that our individual-level mixture models may provide a useful complement to these approaches in such contexts.

6 Appendix A - Derivation of equations (15) and (16)

1. $p(N|\alpha,\lambda)$ (Eq. (15))

$$\begin{aligned}
p(N|\alpha,\lambda) &= \int p(N,z_2,g_2|\alpha,\lambda)d(g_2,z_2) \\
&\propto p(N)\binom{N}{n}\int\prod_{i=n+1}^{N} p\left(g_i|\alpha\right)\prod_{j=1}^{J}\prod_{k=1}^{K}\left(g_{ik}\left(1-\lambda_{jk}\right)\right)^{z_{ijk}}d(g_2,z_2) \\
&= \binom{N}{n}p(N)\prod_{i=n+1}^{N}\sum_{z\in Z}\int_{\Delta}p(g_i|\alpha)\prod_{j=1}^{J}\left(\prod_{k=1}^{K}g_{ik}^{z_{jk}}\right)\left(1-\lambda_{jz_j}\right)dg_i \\
&= \binom{N}{n}p(N)\left(\sum_{z\in Z}\int_{\Delta}p(g|\alpha)\prod_{j=1}^{J}\left(\prod_{k=1}^{K}g_{k}^{z_{jk}}\right)\left(1-\lambda_{jz_j}\right)dg\right)^{N-n} \\
&= \binom{N}{n}p(N)\left(\frac{\Gamma(\alpha_+)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\sum_{z\in Z}\int_{\Delta}\left(\prod_{k=1}^{K}g_{k}^{\alpha_k+z_{+k}-1}\right)dg\prod_{j=1}^{J}\left(1-\lambda_{jz_j}\right)\right)^{N-n}
\end{aligned}$$

where $Z=\{1,2,...,K\}^J$, $z_{jk}=I_{\{z_j=k\}}$, Δ is the $K-1$ dimensional probability simplex and $z_{+k}=\sum_j z_{jk}$. Solving the integral, we get

$$\begin{aligned}
p(N|\alpha,\lambda) &\propto \binom{N}{n}p(N)\left[\frac{\Gamma(\alpha_+)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\sum_{z\in Z}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+z_{+k})}{\Gamma(\alpha_++z_{++})}\prod_{j=1}^{J}\left(1-\lambda_{jz_j}\right)\right]^{N-n} \\
&= \binom{N}{n}p(N)\left[\frac{\Gamma(\alpha_+)}{\Gamma(\alpha_++J)}\sum_{z\in Z}\left(\prod_{k=1}^{K}\frac{\Gamma(\alpha_k+z_{+k})}{\Gamma(\alpha_k)}\right)\prod_{j=1}^{J}\left(1-\lambda_{jz_j}\right)\right]^{N-n}
\end{aligned}$$

2. $p(z_2|N,\alpha,\lambda)$ (Eq. (16))

$$\begin{aligned}
p(z_2|N,\alpha,\lambda) &= \int p(z_2,g_2|N,\alpha,\lambda)dg_2 \\
&\propto \int \left(\prod_{i=n+1}^{N} p(g_i|\alpha) \prod_{j=1}^{J} \prod_{k=1}^{K} \left(g_{ik}\left(1-\lambda_{jk}\right)\right)^{z_{ijk}} \right) dg_1 \\
&= \prod_{i=n+1}^{N} \int_{\Delta} p(g|\alpha) \prod_{j=1}^{J} \prod_{k=1}^{K} \left(g_k\left(1-\lambda_{jk}\right)\right)^{z_{ijk}} dg \\
&= \prod_{i=n+1}^{N} \left[\left(\prod_{j=1}^{J} \prod_{k=1}^{K} \left(1-\lambda_{jk}\right)^{z_{ijk}} \right) \int_{\Delta} p(g|\alpha) \prod_{k=1}^{K} g_k^{z_{ijk}} dg \right] \\
&= \prod_{i=n+1}^{N} \left[\left(\prod_{j=1}^{J} \prod_{k=1}^{K} \left(1-\lambda_{jk}\right)^{z_{ijk}} \right) \int_{\Delta} \prod_{k=1}^{K} g_k^{a_k+z_{ijk}-1} dg \right] \\
&= \prod_{i=n+1}^{N} \left[\left(\prod_{j=1}^{J} \prod_{k=1}^{K} \left(1-\lambda_{jk}\right)^{z_{ijk}} \right) \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_k+z_{i+k}\right)}{\Gamma\left(\alpha_+ + J\right)} \right] \\
&\propto \prod_{i=n+1}^{N} \left[\left(\prod_{k=1}^{K} \Gamma\left(\alpha_k+z_{i+k}\right) \right) \prod_{j=1}^{J} \left(1-\lambda_{jz_{ij}}\right) \right]
\end{aligned}$$

References

Agresti, A. (1994), "Simple capture-recapture models permitting unequal catchability and variable sampling effort," *Biometrics*, **50**, 494–500.

Ball, P., Asher, J., Sulmont, D., and Manrique, D. (2003), "How many peruvians have died. An estimate of the total number of victims killed or disappeared in the armed internal conflict between 1980 and 2000," AAAS. Report to the Peruvian Truth and Reconciliation Commission (CVR). Also published as Anexo 2 (Anexo Estadístico) of CVR Report.

Basu, S. and Ebrahimi, N. (2001), "Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence," *Biometrika*, **88**, 269–279.

Bishop, Y., Fienberg, S., and Holland, P. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press, reprinted in 2007 by Springer-Verlag, New York.

Bruno, G., LaPorte, R., Merletti, F., Biggeri, A., McCarty, D., and Pagano, G. (1994), "National diabetes programs. Application of capture-recapture to "count" diabetes," *Diabetes Care*, **17**, 548.

Comisión de la Verdad y Reconciliación (2003), *Informe Final*, Lima, Perú: CVR.

Cormack, R. (1968), "The statistics of capture-recapture methods," *Oceanographic and Marine Biology Annual Review*, **6**, 455–501.

Darroch, J., Fienberg, S., Glonek, G., and Junker, B. (1993), "A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability," *Journal of the American Statistical Assocation*, **88**, 1137–1148.

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, London: Chapman & Hall/CRC.

Erosheva, E. (2002), "Grade of membership and latent structures with application to disability survey data," Ph.D. thesis, Department of Statistics Carnegie Mellon University.

Erosheva, E., Fienberg, S., and Joutard, C. (2007), "Describing disability through individual-level mixture models for multivariate binary data," *Annals of Applied Statistics*, **1**, 502–537.

Fienberg, S. (1972), "The Multiple recapture census for closed populations and incomplete 2ˆk contingency tables," *Biometrika*, **59**, 591–603.

Fienberg, S., Johnson, M., and Junker, B. (1999), "Classical multilevel and Bayesian approaches to population size estimation using multiple lists," *Journal of the Royal Statistical Society. Series A*, **162**, 383–406.

Goudie, I. and Goudie, M. (2007), "Who captures the marks for the Petersen estimator?" *Journal of the Royal Statistical Society. Series A*, **170**, 825–839.

Holland, P. and Rosenbaum, P. (1986), "Conditional association and unidimensionality in monotone latent variable models," *Annals of Statistics*, **14**, 1523–1543.

International Working Group for Disease Monitoring and Forecasting (1995a), "Capture-recapture and multiple-record systems estimation I: History and theoretical development," *American Journal of Epidemiology*, **142**, 1047–1058.

— (1995b), "Capture-recapture and multiple-record systems estimation II: Applications in human diseases," *American Journal of Epidemiology*, **142**, 1059–1068.

Norris, J. and Pollock, K. (1996), "Nonparametric MLE under two closed capture-recapture models with heterogeneity," *Biometrics*, **52**, 639–649.

Petersen, C. (1896), "The yearly immigration of young plaice into the Limfjord from the German Sea," *Report of the Danish Biological Station*, **6**, 1–48.

Pledger, S. (2000), "Unified maximum likelihood estimates for closed capture-recapture models using mix-

tures," *Biometrics*, **56**, 434–442.

— (2005), "The performance of mixture models in heterogeneous closed population capture-recapture," *Biometrics*, **61**, 868–876.

Pledger, S., Pollock, K., and Norris, J. (2003), "Open capture-recapture models with heterogeneity: I. Cormack-Jolly-Seber Model," *Biometrics*, **59**, 786–794.

Rasch, G. (1980), *Probabilistic Models for Some Intelligence and Attainment Tests*, Chicago: University of Chicago Press, expanded edition of the 1960 work, with forward and afterward by B.D. Wright.

Robert, C. and Casella, G. (2004), *Monte Carlo Statistical Methods*, New York: Springer-Verlag.

Sanathanan, L. (1972), "Models and estimation methods in visual scanning experiments," *Technometrics*, **14**, 813–829.

— (1973), "A comparison of some models in visual scanning experiments," *Technometrics*, **15**, 67–78.

Wittes, J., Colton, T., and Sidel, V. (1974), "Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources." *Journal of Chronic Diseases*, **27**, 25–36.

Woodbury, M., Clive, J., and Garson Jr, A. (1978), "Mathematical typology: A grade of membership technique for obtaining disease definition." *Computers in Biomedical Research*, **11**, 277–98.